даних за допомогою різних фреймворків та алгоритмів, розсилання підписникам новин у науковому світі. З'явиться можливість знайти всі наукові конференції в одному місці та не пропустити цікаві заходи у розсилці.

Література

1. Биков В. Ю. Сучасні завдання інформатизації освіти / В. Ю. Биков // Інформаційні технології і засоби навчання. – 2010. – № 1(15). – Режим доступу до журн. : http://www.ime.edu-ua.net/em.html
2. Закон України "Про Основні засади розвитку інформаційного суспільства в Україні на 2007-2015 роки" (Відомості Верховної Ради України, 2007. – № 12, ст. 102) [Електронний ресурс]. – Режим доступу : http://zakon2.rada.gov.ua/laws/show/537-16. – Загол. з екрану.
3. Дефіцит ІТ-фахівців в Україні становить 30% // УНІАН ОСВІТА, 29.04.2011 [Електронний ресурс]. – Загол. з екрану. – Режим доступу : http://education.unian.net/ukr/detail/190597

РЕЗАНОВА В.Г., НЕСІН К.О.

ПРОГРАМНІ ЗАСОБИ ДЛЯ АВТОМАТИЗОВАНОГО АНАЛІЗУ ТА ПОПОВНЕННЯ КОНТЕНТУ

REZANOVA V. G., NESIN K.A.

SOFTWARE FOR AUTOMATED ANALYSIS AND REFRIGERATION OF THE CONTENT

The purpose of work is to create a research laboratory with the function of automatic content replenishment. The content must be pre-analyzed using an algorithm that will ensure the quality of the selected material.

Thousands of sites disappear everyday on the World Wide Web, some of them contain useful information that can be restored.

To implement task used .NET technology, ASP.NET MVC, MS SQL and Html Agility Pack, should also develop my own algorithms replenishment content.

The work is based on modern web technologies and knowledge about the principle of the Internet, which provides the automatic replenishment of the content of the research laboratory on the basic information from the removed sites.

Keywords: ASP NET MVC, automatic refill of the content, web-page content identification.

Вступ

Інформація являється найціннішим ресурсом в сучасному світі. Саме тому питання відновлення втраченого в мережі Інтернет контенту настільки актуальне.

Збереженням інформацій займаються всесвітні архівні служби, але для того, щоб цей контент був доступний користувачам, необхідно навчитися правильно виділяти текст з html сторінки сайтів. Для цього використовуються різноманітні алгоритми вилучення тексту, що відрізняються по якості отриманого результату та швидкості роботи. [4].

Основне завдання полягає в створені науково-дослідної лабораторії, яка буде автоматично поповнюватись контентом, що зник з пошукового

індексу, розподіляти його по різним категоріям, в залежності від тематики текстів.

Реалізація даного завдання потребує використання алгоритму вилучення тексту, продуманої логіки розподілу отриманого контенту.

Основна частина

Дослідження теоретичної частини та аналіз попередніх експериментів показав, що щоденно можна аналізувати близько 5000 джерел інформації, серед яких будуть десятки, що відповідають тематиці нашої науково-дослідної лабораторії.

Основна частина роботи полягає в створенні алгоритму виділення тексту, умовно його можна представити в наступному списку дій:

- отримання html коду сторінки, яка нас цікавить;
- попередня перевірка сторінки (чи текст сторінки включає слова російської або української мови, підрахунок кількості слів);
- виділення фрагменту сторінки, який включає текст;
- очищення отриманого фрагменту від зайвої розмітки та атрибутів html-тегів;
- повторна перевірка отриманого результату на відповідність фрагменту до тексту.

Виділення фрагменту сторінки, який включає контент, являється основною проблемою в даному алгоритмі, яку можна вирішити різними способами. Варто однозначно відрізняти звичайний фрагмент тексту від цілісної статті. [1][2]

Серед первинних критеріїв, що відрізняють тематичну статтю від будь-якого іншого текстового фрагменту, варто звернути увагу на:

- довжину текстового фрагменту;
- кількість абзаців в html-структурі та довжина кожного з них, цей параметр являється важливою характеристикою статті;
- наявність обов’язкових елементів тексту: ком, крапок. Взагалі пунктуація – цінний покажчик, без врахування якого пошук контенту приносив би набагато менш якісні результати;
- за тематичність тексту відповідає наявність ключів пошуку (ми можемо реалізувати декілька варіацій пошуку по ключам).

Серед основних помилок з якими можливо зіткнутись при виділенні тексту варто звернути увагу на наступні:

- пошук ключів серед атрибутів html-тегів;
- ігнорування довжини текстових фрагментів в кожному окремому тегу;
- ігнорування необхідності видалення рекламних блоків (або інших нетематичних), які можуть зустрічатися в тексті статті на будь-якому сайті;

- ігнорування необхідності видалення останніх фрагментів тексту в статті, якщо вони належать до тегів, нетипових для текстової html-розмітки на конкретній сторінці.

Остання помилка часто являється причиною появи в виділеному тексті мета-інформації (автор, дата, написання, посилання на джерело), списку посилань на схожі матеріали, коментарів.

Іншим не менш важливим завданням є підготовка теоретичної основи: визначення основних категорій науково-дослідної лабораторії, ключових слів, які допоможуть визначати тематику контенту, алгоритм та джерела відбору текстів. Потрібно порівняти різні джерела та методики відбору, щоб забезпечити найкращий результат.

Ми збираємо текстові матеріали, що об'єднані однією тематикою, але для зручності користувачу необхідно відобразити їх за допомогою зрозумілого інтерфейсу. Саме для цього потрібно створити науково-дослідну лабораторію з розподілом контенту на різноманітні категорії.

Науково-дослідна лабораторія буде створена за допомогою технології ASP.NET MVC, що дозволить зробити її масштабованою та зручною в використанні.

Так як ми плануємо використовувати контент с html-сторінок на своїй науково-дослідній лабораторії, нам потрібно зберегти його логічну цілісність та структуру. Тому під час обробки сторінки ми будемо використовувати Html Agility Pack. Ця бібліотека допоможе виділяти html фрагмент, що містить текст, та очищати його від зайвих елементів. [5].

Для роботи нашого сервісу потрібно щоденно перевіряти бази видалених доменів на появу нових даних. Очікувана кількість щоденних нових надходжень складає близько 6 тисяч доменів. Після цього нам необхідно перевірити кожний на наявність архівних копій його сторінок. Для цього будуть використані публічні архіви мережі Інтернет.

Далі варто відсортувати отримані сторінки, проаналізувавши стартову сторінку кожного сайту, ми можемо використати попередньо визначені ключі для перевірки тематики джерела. Після сортування у нас залишаться лише ті джерела, що відповідають нашій тематиці. Зібравши ці вхідні дані варто перейти до обробки кожної сторінки, використовуючи наш алгоритм вилучення тексту. [3][6] Цю частину роботи можна виконати в якості консольного додатку, адже користувачу не потрібно задавати ніяких вхідних даних.

Наступним етапом розробки стане створення веб-інтерфейсу за допомогою ASP.NET MVC, який буде відображати всі отримані результати та візуально сортувати їх для кінцевих користувачів в вигляді різноманітних категорій сайту. Для поєднання окремих частин в одну систему будемо використовувати роботу з базами даних MS SQL.

Збереження інформації в базі даних буде доцільно і для систематизування отриманого контенту. Для якісної роботи системи варто виключити ситуації повторної перевірки одного і того ж джерела текстів. Також база даних стане джерелом вхідних даних, ключів пошуку.

Висновки

Розробка програмного забезпечення, що реалізує всі вищеописані кроки, дозволить раціоналізувати роботу з втраченим контентом. З'явиться можливість зручно переглядати знайдений контент, користуючись науково-дослідною лабораторією.

Важливу частину роботи займає алгоритм виділення тексту, який може бути використаний в подальшому для різноманітних цілей в сфері обробки текстових даних.

Література

1. J. Gibson, B. Wellner, and S. Lubar. Adaptive web-page content identification.
2. Hung-Yu Kao, Jan-Ming Ho. WISDOM: Web Intrapage Informative Structure Mining Based on Document Object Model.
3. Christian Kohlschütter, Peter Fankhauser, Wolfgang Nejdl. Boilerplate Detection using Shallow Text Features.
4. http://www.interface.ru/home.asp?artId=27160.
5. http://html-agility-pack.net/api
6. https://habr.com/post/99918/.

РЕЗАНОВА В.Г., ЩЕПАНКОВСЬКИЙ С.А.

ПРОГРАМНЕ ЗАБЕЗПЕЧЕННЯ ДЛЯ СТАТИСТИЧНОЇ ОБРОБКИ РЕЗУЛЬТАТІВ АНТРОПОМЕТРИЧНИХ ВИМІРІВ

REZANOVA V.G., SHCHEPANKOVSKIY S.A.

SOFTWARE FOR STATISTICAL PROCESSING OF ANTROPOMETRIC MEASUREMENTS RESULTS

The need to improve the design of the indoor footwear and the introduction of new automated design systems is the basis for maintaining the competitiveness of domestic footwear. At present, a large number of experimental results are accumulated, manual processing of which is labor-intensive and long-lasting. The purpose of the work: the development of software for processing experimental data of anthropometric measurements of a person in order to ensure the possibility of further building a rational form of shoes shoes with the use of high-tech equipment.

Practical value. The developed software provides the possibility of integrated design of shoe shoe shoes, taking into account technological and design features and ergonomic requirements to the shape of the body of the pad, which will increase the effectiveness of design processes and reduce their duration.

Elements of scientific novelty. For the first time a special software application was created for statistical analysis of anthropometric parameters and plotting patterns of the normal distribution of dimensional features.

Keywords: anthropometric measurings, statistical treatment, software.